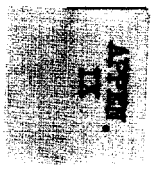
Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7



Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

APPENDIX IX

Files and Procedures for Code Development

The development of an indexing system to exploit modern electronic equipment is an intensely practical problem. Although theoretical considerations can contribute much of importance, the experimental approach must be taken if development of a machine indexing system of optimum effectiveness is to be achieved within a reasonable time.

In order to expedite development of our indexing and coding system, two somewhat different approaches were pursued simultaneously in processing individual terms and working out codes for them. One approach was based on the more general terminology of the Intelligence Subject Code. Another involved a much larger number of terms collected from subject heading lists, classification systems, and indexes of textbooks in the general field of science and technology. In both approaches, semantic factoring was the basis for code construction.

A record file used to analyze ISC terminology was built up by punching each line--that is each heading and sub-heading of the ISC-- in a separate IBM card. The new scanning machine was used to search the file for each term, e.g. "radiation". The groups of cards having some term in common were then run through the automatic typewriter and the various lines of the ISC containing the term in question typed out on a separate card. In this way a file was built up in which each term is followed by cards showing how it is used in the various headings of the ISC. The file resembles a concordance or "index verborum" commonly used in theological studies.

A second file, in which ISC terms were included embraced a much wider range of terms. It was anticipated, at the time of initiating our general development program, that the terminology of the ISC would be particularly appropriate as building blocks for constructing the general framework of the new indexing system. It was expected, however, that additional terminology would also be needed to provide the discriminating power desired to meet fully the requirements of OSI, in particular, and of the Agency, in general. It was decided to limit our efforts with regard to processing additional terminology to the broad field of science and technology. The first step was to collect appropriate terms and to this end a number of libraries, editorial offices and information centers were visited and a collection was made of subject heading lists, classification systems and indexes of general textbooks in the broad field being covered. Details

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

concerning this collecting operation and a list of the various terminology sources which were processed are provided in Appendix VIII. This collecting effort was directed to bringing together terminology which the experience of others had shown to be particularly effective in indexing and classifying scientific and technical information. All terms found in the ISC were included in this file.

It was realized at the start that processing of terminology to establish an indexing system can be expedited by following a definite plan and using appropriate mechanical aids. Figure I entitled "Terminology Flowchart" presents schematically the various processing steps.

The starting point for processing terminology is obtaining an understanding of meaning and usage. The first step in the processing operation was to construct a hand-sorted punched card file, with a single term entered on each card together with its dictionary definition. (See Figure II.) Thanks to the cooperation of the G & C Merriam Co. publishers of "Webster's International Dictionary", two unbound copies of the latest unabridged edition were made available. Definitions of terms entered in our file were clipped from the unbound pages and the clippings attached to the punched cards. It was realized, of course, that the Webster dictionary can scarcely hope to keep up to date with the continuing evolution and development of terminology, particularly in those fields in which progress is being made at a rapid rate. Consequently, additional notes as to usage of terms were provided by consulting special dictionaries in such fields as electronics.

Once definitions and, where necessary and feasible, further notes on meanings had been entered on the hand-sorted punched cards, the terms were categorized in a rough preliminary fashion in two different ways. One form of categorization typified the term with regard to its semantic content under the broad headings of: 1. Processes, 2. Machines, apparatus, devices, 3. Materials, substances, 4. Attributes, characteristics, 5. Abstract concepts. When categorizing terms under five general headings given above, certain rules and procedure were set up as follows. For example, the category "Machines" was considered to embrace mechanisms and related devices without regard to materials from which they had been constructed. The fact that a certain term falls within a given category was indicated by appropriately punching the card on which the term and its definition had been entered.

The second way in which terminology in this file was categorized concerned the general field in which a given term is used. Here again five main catagories were set up. These were: 1. Chemistry. 2. Physics 3. Mechanics 4. Biology. 5. General.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7
Security Information

The fact that a given term is used within one or possibly several of these broad fields was indicated by again appropriately punching the card on which the term and its definition had been entered.

Categorization of terms under general headings effects a useful breakdown for further analysis, even though the assignment of a given term to one category or another is not in certain cases straightforward and obvious. Assignment of certain terms to more than one general heading appears desirable to avoid excessive arbitrariness. A file of about 30,000 terms was processed in this manner.

In general the categorizing of terms from the two viewpoints mentioned above suffices to group together similar terms to facilitate further study. Terms categorized both under "Machines, apparatus, devices" and under "Chemistry" constitute a group having a number of similar basic characteristics.

General categorization served as a first step for further classification in a more detailed scheme which was undertaken for two purposes. 1. to provide smaller groupings of more closely related terms so as to facilitate studying relationships between them and 2. to provide lists which can be examined by specialists in the various fields. (see Figure IV).

It will be noted that this classification scheme does not resemble an ordinary classification system. This is because the purpose of the scheme was to group terms to conform to existing technical specialties and thus facilitate consideration of terms falling within the same general category. In order to make it possible to prepare lists of classified terms conveniently, an IBM file was prepared in which the punching spelled out each term and also indicated its classification. Subsequently the punching was extended to indicate the semantic factoring of the various terms involved.

As already noted the semantic factoring of terminology is an operation never previously undertak: for the purpose of developing an indexing and searching system. Perhaps the closest approach was the development of basic English. Before undertaking to develop and apply the semantic factoring technique to our terminology file, consideration was given to the words constituting basic English. It was observed that these words apply mostly to everyday situations and as a consequence, could not be applied directly with advantage to the problem of developing an indexing and coding system for machine searching of scientific and technical information. Consideration of basic English did, however, provide general background useful in developing semantic factoring as a technique in code development.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7
Security Information

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7
CONFIDENTIAL
Security Information

It might have been possible, when initiating semantic factoring to prepare an a priori list of factors on a theoretical basis. This seemed inadvisable. Instead the following approach was taken. The three people who undertook to develop the semantic factoring technique first discussed its purpose and arrived at agreement as to how semantic factoring was to be developed as an essential element in our indexing and encoding system for searching by machine. Each of the three people then took a small group of terms--approximately 200 each--carefully considered them and related them to those more general concepts which appeared useful and appropriate for facilitating machine search operations. Discussion of this preliminary work revealed a considerable area of agreement but also certain differences of opinion which were then discussed both for the purpose of arriving at decisions whenever possible, and also for defining problems requiring further consideration. Particular attention was directed to preparing and checking for consistency a preliminary list of semantic factors. These factors were then applied to approximately 2,000 terms and the initial list of factors supplemented where it seemed appropriate. The new list was re-examined as to consistency and probable eventual usefulness in expediting machine searching.

Three persons having scientific and technical background then proceeded with the task of semantic factoring an other terms in the file. The factors were encoded for each term and the code representation of the factors punched in the IBM cards for the individual terms. (See Figure III.) The cards were then sorted out on the basis of the individual factors and preliminary lists run off on a tabulating machine showing the terms to which each factor had been applied. Similar lists were prepared for terms according to the field in which the term is used. These two sets of lists were used to check the consistency of the semantic factoring and also to re-evaluate the eventual usefulness of individual semantic factors for the purpose of providing reference points for machine searching. In reviewing the provisional semantic factors, it has been observed that a few of them, for example the concept "product", might serve more effectively as a role indicator (See Appendix X) which would be attached, as appropriate, to the code for any given entity. Similarly certain concepts, e.g. "agriculture", might perhaps be more advantageously incorporated in a list of terms used to indicate the general field (or fields) to which a given document pertains.

This review of the preliminary lists was followed by extensive revision of the IBM file used to generate it. The revised IBM file has been used to prepare listings of terms on the basis of their semantic factors. See Appendix XI for a detailed description of these lists and their use for encoding index entries.

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7
CONFIDENTIAL
Security Information

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

TERMINOLOGY FLOWCHART

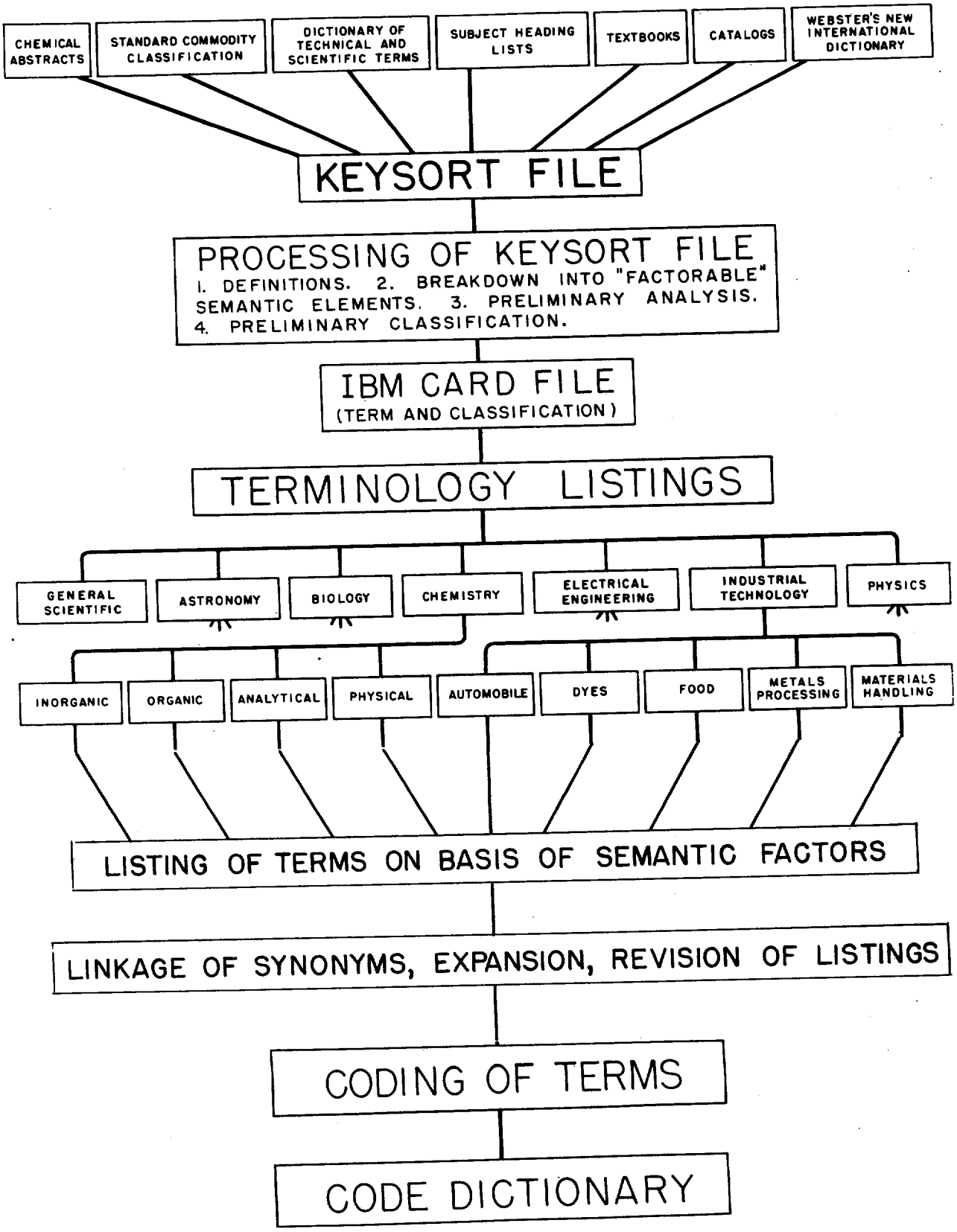


Figure 1

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

Preliminary processing of terms on Keysort cards

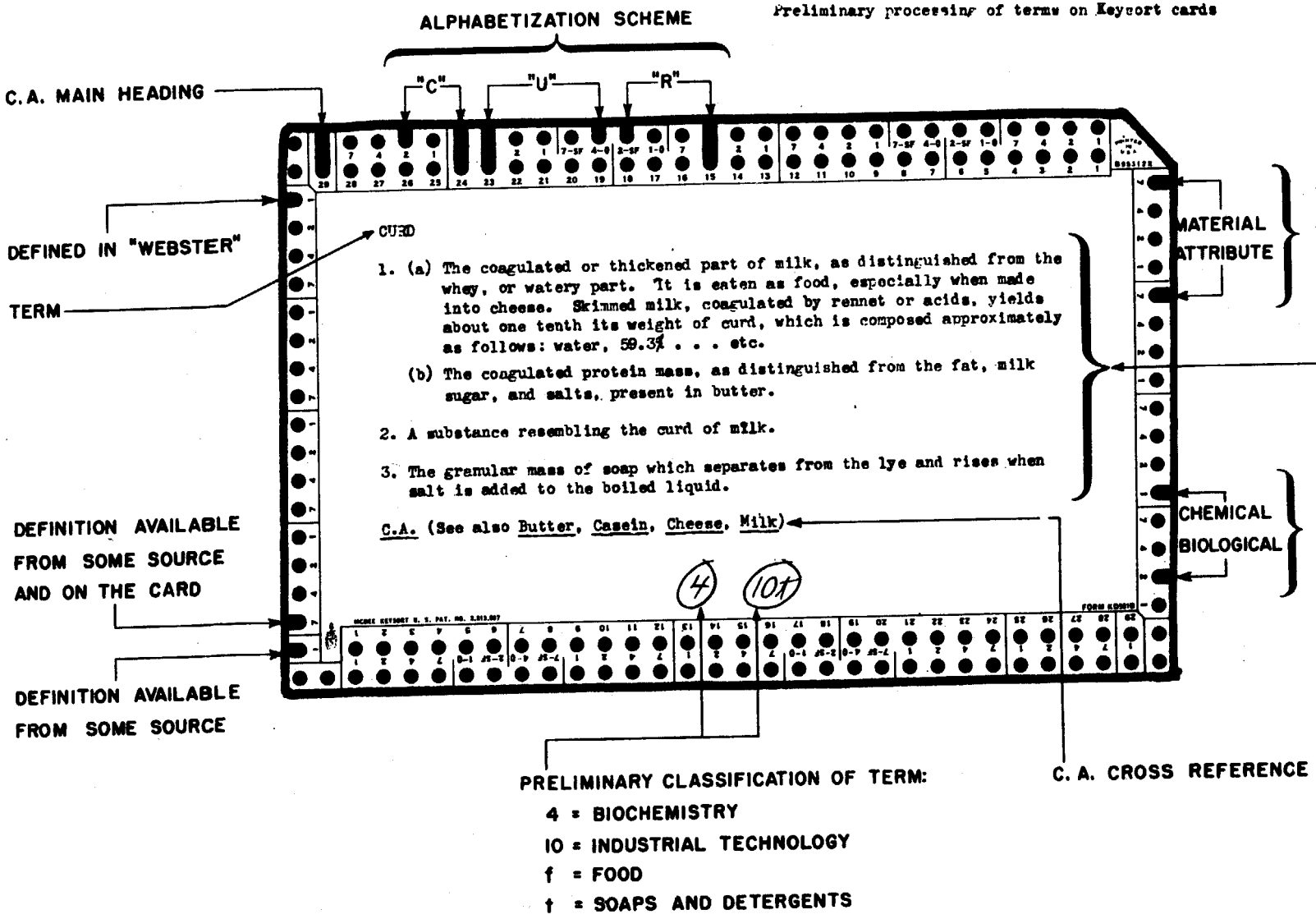


Figure 2

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

Transfer of information to IBM card for automatic manipulations

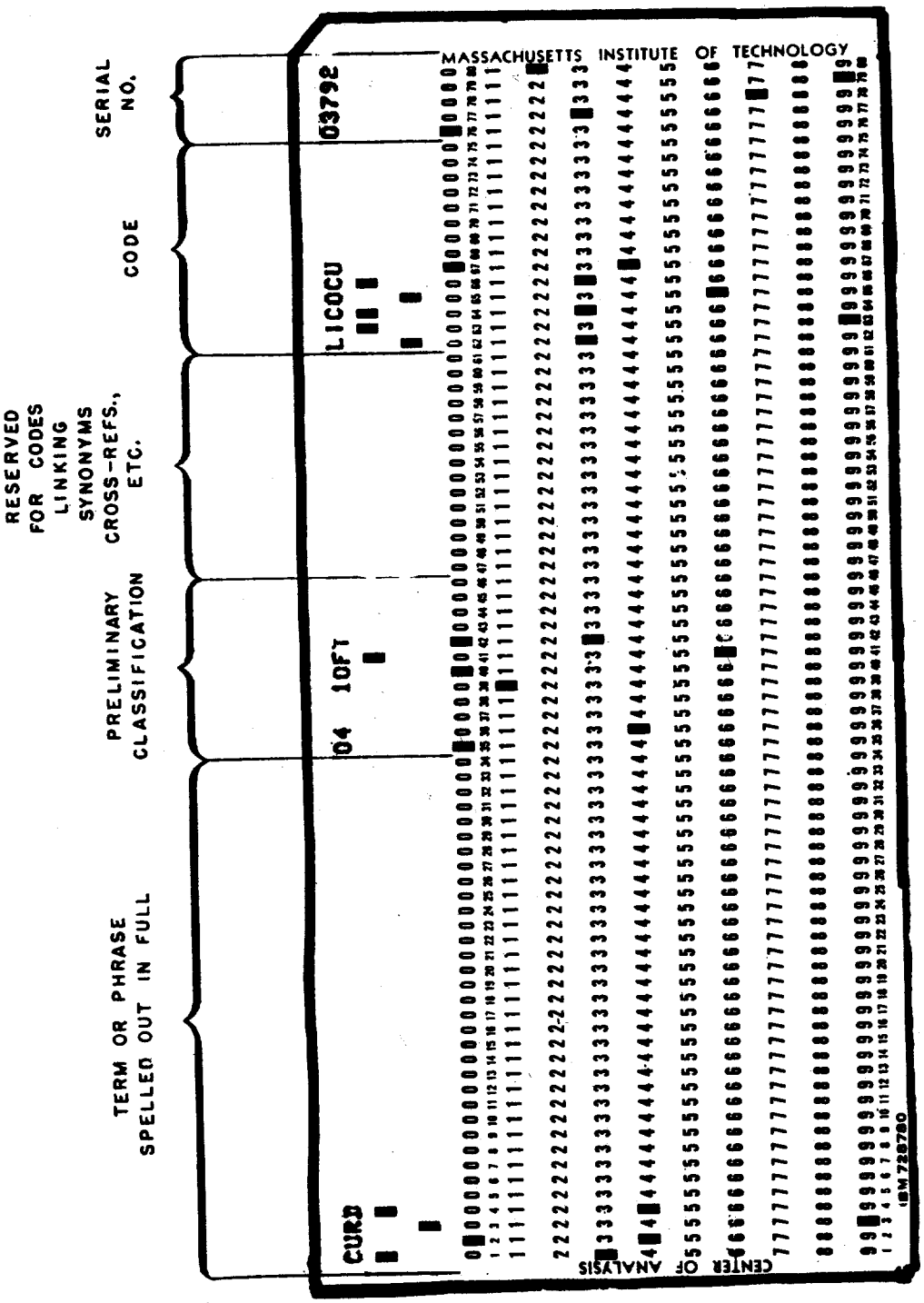


Figure 3

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7
Security Information

FIGURE IV

Preliminary Classification Scheme

00. General Scientific

01. Acoustics

02. Astronomy
    Astrophysics
    Meteorology
    Navigation

03. g. General Biology (/ terms with meanings in zoology and botany which are not the same)

    z. General Zoology
    P. Animal parts and organs

    b. Botany

    a. Agriculture and plant pathology

04. Biochemistry

    g. general
    r. reagents and materials (incl. classes such as anticoagulants)
    t. tests and reactions

05. Chemistry
    G = General
    I = Inorganic
    O = Organic
    A = Analytical
    P = Physical and Theoretical
    L = Apparatus

07. Civil Engineering
    Architecture and Building Trade
    Sanitary Engineering
    Structural Engineering (mass structures)
    Surveying (general)

08. Electrical Engineering
    p. Power
    c. Communication
    i. Illumination

09. Electronics

10. Geology and Mineralogy (including Geography)

11. Industrial Technology
    a. General Industrial
    b. Automotive
    c. Dyes and Textiles
    d. Explosives
    e. Fertilizers
    f. Food and Fermentation
    g. Glass, Ceramics, Cement
    h. Leather
    i. Materials handling
    j. Metals Processing
    k. Paint, Varnish, Lacquer
    l. Pesticides
    m. Petroleum & Solid Fuels
    n. Pharmaceuticals, Cosmetics, Perfumes
    o. Photography
    p. Plastics
    q. Printing
    r. Pulp and Paper
    s. Rubber
    t. Soaps and Detergents
    u. Wood Industries
    w. Fur Industries
    x. Tobacco Industries
    z. Electrical Appliances

12. Mathematics

Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150014-7
Security Information

Approved For Release 2000/[redacted]P57-00042A000200150014-7
Security Information

13. Medicine
    - g. general
    - d. diseases and abnormalities
    - i. instruments and devices
    - o. operations and procedures
    - v. veterinary

14. Mechanical Engineering
    - Hydraulics
    - Air Conditioning, Refrigeration, Heating
    - Materials Study (strength, etc.)
    - Tools and Mechanical Devices

15. Military Engineering and Ordnance

16. Mining Engineering

17. Nuclear Science

18. Physics
    - G. General
    - E. Electricity and Magnetism
    - O. Optics and Light
    - M. Mechanics
    - T. Theoretical Physics
    - H. Heat
    - W. Weights and Measures
    - P. Physical property of matter

19. Transportation Engineering
    - Air
    - Sea
    - Land

Approved For Release 2000/08/[redacted]57-00042A000200150014-7
Security Information